1К

УДК 519.683.5:004.934:81'367

Ю.И. Кисленко, А.В. Терентьев
Национальный технический университет Украины
«Киевский политехнический институт», г. Киев, Украина
bary@bigmir.net

Проблемы и перспективы развития поисковых систем

Рассмотрены важнейшие проблемы функционирования ведущих поисковых систем – недостаточный учёт и использование естественного языка на этапах формирования запроса и поиска информации. Перспективы видятся в использовании нового подхода к структурной организации языка, снимающего много противоречий классической лингвистики и позволяющего заменить «поиск по ключевым словам» на «поиск по знаниям».

Текущее состояние поиска в INTERNET

Поиск как один из аспектов моделирования речевой деятельности. Поиск в INTERNET – одна из наиболее продвинутых и воспринятых в мире современных информационных технологий, значительно облегчившая и ускорившая коммуникативные процессы, связанные с хранением, накоплением и использованием разноплановой (текстовой, графической, акустической и др.) информации, гарантируя при этом сохранность качества документов на момент цифровой обработки. На электронные носители перенесено огромное количество информации. Автор был свидетелем, когда в конце девяностых годов еще только ставились вопросы компьютеризации языков (Димитар Шишков – пионер болгарской компьютерной школы и неутомимый энтузиаст переложения на электронные носители всей информации, накопленной человечеством [1]). Многие, только еще поднимавшиеся тогда вопросы, сегодня решены; однако ряд проблем представления, накопления, хранения и использования знаний в естественно-языковой форме все еще актуальны и остаются нерешенными по сей день. Далее пойдет речь о текущем состоянии информационного поиска, о вопросах представления и использования знаний, хранимых на электронных носителях. Предварительно необходимо представить платформу, с которой рассматриваются задекларированные вопросы.

Информационный поиск представляет собой попытку моделирования лишь одной из сторон многогранной *речевой деятельности* (РД) человека. Получив в руки мощное средство представления и обработки символической реальности [2], заданной в текстовом или речевом виде и сформированной в качестве глобальной компьютерной сети, человечество (в который уж раз!) обольстилось надеждой успешного моделирования интеллектуальной деятельности человека, связанной с вопросами хранения, накопления, структурирования, поиска и понимания символической информации. В работе ограничимся рассмотрением лишь вопросов обработки естественно-языковой информации, представленной на текстовом или речевом уровнях.

Конечно, информационный поиск представляет собой лишь одну из сторон многогранной речевой деятельности человека, связанной с хранением, обработкой и поиском необходимой информации. Речевая деятельность, в общем случае, представляет нам еще целый спектр направлений, которые в той или иной мере отражены в современных информационных технологиях. Перечислим важнейшие направления моделирования речевой деятельности человека, которые уже в определенной степени реализованы в современных информационных технологиях:

поиск в INTERNET – моделирует способность человека обращаться к накопленным знаниям, учитывая всевозможные схемы трансформирования и преобразования информации; *экспертные системы* – моделируют процессы накопления и обработки знаний в определенных достаточно узких предметных областях с целью формирования и принятия решений в конкретных условиях и ситуациях; *синтез/анализ текстовой информации* – практически обязательная процедура всех технологий, связанных с обработкой текстовой информации; *синтез/анализ речевой информации* – мощное направление, связанное с вопросами обработки речевой деятельности человека; *автоматический перевод* – моделирование многоязычной компетенции человека с целью трансляции речевой/текстовой информации с одного языка на другие; *естественно-языковой интерфейс* – технологии, моделирующие возможности общения с компьютером посредством голосового тракта, а не через клавиатуру (Билл Гейтс грозился через десять лет выбросить клавиатуру – не вышло); – *автоматическое формирование, накопление и использование знаний.*

Перечисленные аспекты речевой деятельности в комплексе представляют интеллектуальную деятельность человека, отдельные стороны которой мы пытаемся воплотить в ряде современных информационных технологий. Еще раз подчеркнем, что указанные направления формируют целый кластер информационных технологий, для которых общей платформой будет естественно-языковая форма представления информации. Поэтому все названные технологии обозначим единым термином *информационные естественно-языковые технологии* (ИЕЯТ). Этот термин адекватно отображает важнейшую особенность указанных технологий и давно используется автором; особенность в том, что во всех этих случаях общей проблемой выступает обработка естественно-языковой информации. Состояние указанных ЕЯ-технологий детально проанализировано в работах автора [3], [4]; здесь мы кратко проанализируем важнейшие аспекты текущего состояния лишь поисковых систем, ссылаясь на признанные авторитеты.

Относительно возможностей и качества поисковых систем сегодня уже формируется определенный скепсис – глобальная система знаний пока еще не может быть осмыслена в структурном и функциональном планах небольшим интеллектуальным коллективом разработчиков. Первоначальная эйфория как пользователей, так и идеологов постепенно сменяется непредубежденным анализом реальных возможностей и перспектив.

Реальное положение поиска в мировой паутине. Здесь ничего не остается как согласиться с компетентным мнением ведущих специалистов в области информационного поиска:

– «Среду World Wide Web можно сравнить с огромной энциклопедией, насчитывающей более 800 млн страниц, которую забыли оснастить оглавлением» (на запрос «INTERNET» откликается более 4 млн страниц); однако жизнь, время и рынок диктуют свои законы – искать более эффективные способы удовлетворения информационной потребности пользователей» [5];

– «INTERNET похож на большую свалку – там есть все, но найти это невозможно» [6];

– порочен сам принцип формирования базы знаний: традиционные Web переполнены неструктурированной информацией [7], в каждой из областей знания создаются миллионы ресурсов, лишь малая часть которых содержит оригинальную информацию и, следовательно, по запросу получаем много тысяч ссылок на различные ресурсы – анализ их занимает много времени, но все равно не позволяет сформировать уверенность в том, что не пропущены самые ценные ресурсы;

– после выполнения процедуры поиска нет никакой гарантии, что в INTERNET-ресурсах не осталось нужной нам информации;

– основные проблемы связаны с обработкой естественно-языковой информации;

– нерелевантность поиска (информационный шум из-за многозначности ключевых слов – явление синонимии); тематическое смешение результатов поиска;

– неполнота поиска – вместо поиска информации нам представляется поиск ссылок и вместо мгновенного доступа необходимая информация отдаляется за некоторый барьер;

– вместо поиска в ширину получаем поиск в глубину;

– по запросу «INTERNET» Яндекс, например, предлагает 602 млн страниц;

– желательно бы общаться на естественном языке, но не на уровне ключевых слов;

– вместо «глобальной базы знаний или коллективного разума» создан «глобальный хаос» [8];

– за более чем десятилетний период Web настолько развился, что близок к состоянию переполнения (проклятие многомерности!);

Основные проблемы. Обобщив предыдущие оценки, можем теперь акцентировать внимание на важнейших проблемах развития поисковиков:

– компьютерная программа не способна, загрузив произвольный документ (будь-то Web-страница или какой-то файл), понять его содержание. *Требуется все равно программист, который должен разобраться в них и понять смысл или семантику каждого из тегов.* С точки зрения компьютера сеть WWW – это полная неразбериха; выход – семантический Web [9];

– у компьютера нет надежного средства обрабатывать семантику документа [10]. *Семантическая сеть превращается в модель мира, но по мере ее создания, роста размеров и масштабов она становится быстро неконтролируемой*;

– *нет единственного универсального средства решения проблемы релевантности и полноты* – необходимо разумное комбинирование всех доступных средств [11].

Какими же видятся перспективы построения «всемирной базы знаний» самим разработчикам?

– построение всеохватывающей онтологии;

– создание семантического WEB, ориентированного на работу с полными естественно-языковыми текстами без дополнительной разметки;

– контекстные технологии;

– аннотации RDF – декларирование объектов, атрибутов и отношений между ними;

– система должна знать, *что она знает,* т.е., требуется метазнание о знании;

– необходимо понимание естественного языка;

– должны проводиться тщательные исследования *синтаксиса и семантики.*

1 Перспективы развития поисковиков

Взгляд со стороны. Обобщенный вывод краткого обзора текущего состояния информационного поиска напрашивается следующий: на порядок дня встают проблемы структурирования накопленных знаний, формирования семантического Web и организации общения пользователя с Мировой паутиной посредством естественного языка с использованием диалога. *Практически основная проблема сводится к общению с мировой сетью посредством естественного языка как символической формы отображения произвольной информации.* Весьма показательной в этом плане выступает работа Г.С. Осипова о перспективах построения семантического Web [12], где центральной проблемой выступает возможность использования естественного языка.

Позвольте теперь авторам высказать свое видение решения указанных проблем, тем более что уже имеются для этого весомые аргументы. Информационный поиск представляется лишь одним из аспектов функционирования речевой деятельности человека, которая реализуется индивидуальной речевой системой (ИРС), и по признанию ведущих

специалистов лингвистики, психологии, философии, кибернетики базируется на двух составляющих – индивидуальной языковой компетенции человека (на осознанном, либо чаще подсознательном уровне) и накопленных на текущий момент разноплановых знаниях об окружающем мире – среде его обитания. Термин ИРС был введен Л.В. Щербой [13], а согласно существующим современным тенденциям первую составляющую определим как лингвистический процессор (ЛП), а вторую – как базу знаний (БЗ), где зафиксирована модель среды нашего обитания, а в общем случае – модель внешнего мира. Важно отметить, что ЛП и БЗ находятся в диалектическом единстве, структурно единообразны и работают друг на друга. Индивидуальная речевая система функционирует в режимах синтеза (говорения) или анализа (понимания) текстовой/речевой информации. Все перечисленные ранее современные информационные технологии пытаются моделировать лишь отдельные стороны речевой деятельности человека, не сводя их в единую проблему. Если же мы хотим моделировать более-менее адекватно особенности функционирования речевой системы человека, естественно, встает вопрос о формировании модели ИРС как диалектического единства ЛП и БЗ. А что имеем на сегодняшний день?

Первая составляющая ИРС – лингвистический процессор, определяющая языковую компетенцию индивида, в данный момент представляется существующими достижениями классической лингвистики, спрессованными в грамматиках (фонология, морфология, синтаксис, семантика), всевозможных словарях, руководствах и т.п., т.е. всем потенциалом, накопленным за 350 лет своего существования (с момента выхода в свет в 1660 г. первого квалифицированного исследования языка – Грамматики Пор-Рояля). Вторая же составляющая – база знаний, где интегрируются все наши знания о себе и среде нашего обитания, практически остается белым пятном в нашем языкознании со всеми нерешенными проблемами (что такое «знание», как идет процесс его накопления, как организовано взаимодействие с ЛП и т.д., и т.п.).

Вывод кардинальный: если мы хотим более-менее адекватно моделировать отдельные аспекты речевой деятельности человека, необходимо построить модель нашей лингвистической компетенции M_1, создать модель M_2 нашей среды обитания (модель мира) и замкнуть их друг на друга через соответствующий интерфейс. Эта идея, хотя и неявным образом, в свое время озвучена была Г.П. Мельниковым: только при наличии модели мира (среды нашего существования) возможна интерпретация (понимание) речевого сообщения; естественно, здесь сразу возникают вопросы взаимодействия образной и символьной информации [14]. Резюмируя, можем сказать, что каждая из существующих информационных естественно-языковых технологий на сегодняшний день пытается моделировать лишь отдельные аспекты речевой деятельности человека, не затрагивая в целом важнейшие особенности существования и функционирования самого объекта исследования – индивидуальной речевой системы. Это касается, собственно, и существующей идеологии формирования общей концепции информационного поиска – это общая беда всех современных технологий, ориентированных на обработку естественно-языковой информации.

Возможности современной лингвистики. Может ли современная лингвистика удовлетворить запросы поисковиков? К сожалению, приходится констатировать: текущее состояние классической лингвистики не в состоянии обеспечить разработчиков средствами кардинального изменения положения дел. Причины здесь двоякого плана: во-первых, по утверждениям самих же лингвистов высшей квалификации [15], собственно лингвисты внесли лишь незначительный вклад в развитие информационных технологий, во-вторых, на данном этапе лингвистика (снова же классическая) не в состоянии разрешить те проблемы, которые выдвигают нам современные технологии, стремящиеся все полнее моделировать речевую деятельность человека.

В работе «Информационный подход к структурной организации языка» [3] дан тщательный анализ существующего положения дел в классической лингвистике. Очень кратко можно резюмировать: на данный момент наши знания о языке носят спорадический (не системный [16]) характер, в грамматиках больше исключений, нежели правил, не определен сам объект синтаксических исследований, лингвистическая общественность давно созрела к отрицанию существующего деления на простые/сложные предложения [17], не определено до сих пор само понятие словосочетания. Почему это так? – ответ находим у Б.Ю. Городецкого [18]: *«Многие беды в языкознании происходят из-за того, что до сих пор язык считают формой отображения мысли, а не способом организации и представления знаний».* Вот основные причины всех проблем языкознания.

Новый взгляд на структурную организацию языка (новый синтаксис*)*. В противовес классической лингвистике автором Ю.И. Кисленко предложен новый подход к структурной организации языка, который рассматривает текст как конечный продукт интеллектуальной деятельности человека и учитывает при этом давно известные особенности восприятия и обработки информации нашей нервной субстанцией, в частности, зрительным анализатором. Предтечей такого похода стали работы «Нейрофизиологические основания структурной организации языка» в докладах НАН Украины [19] и монография «От мысли к знанию» [20], оказавшиеся фундаментом нового взгляда на синтаксис.

Этот подход позволил на новых основаниях представить структурную организацию языка, основные положения которой представляются следующими:

– с позиций последних достижений нейрофизиологии удалось формально определить понятие *«ситуации»* как элемента восприятия внешнего мира – эта составляющая определяется генетическим уровнем нейронной организации зрительного тракта и не зависит от расы, нации, языка;

– вербализованная форма описания отдельной ситуации определяется как *базовая семантико-синтаксическая структура* (БССС), которая определена на содержательном, графическом и формальном уровнях; это двусоставная монопредикатная структура описания произвольной ситуации внешнего мира, все составляющие которой определены на атрибутивном уровне;

– *структурный уровень организации языка, следовательно, определяется одной-единственной структурой БССС*, а все многообразие структурных форм сообщения вкладывается в монопредикатный или полипредикатный уровни использования БССС;

– *при таком подходе нам не нужна концепция словосочетания*: любые структурные разновидности (кроме идиоматических образований, конечно) определяются как трансформации БССС либо на монопредикатном уровне, либо как схемы их взаимодействия на полипредиктаном уровне;

– в итоге, *структурная организация представляется в виде стройной системы из множества однотипных структур*, и это – справедливо для всех языков;

– очень важно еще указать, что вербализованная форма описания отдельной ситуации в виде БССС на содержательном уровне представляется как отдельный *«квант знаний»* относительно описываемой ситуации; все знание (*фрагмент знания*) представляется в виде взаимосвязанной совокупности отдельных квантов знаний, реализованных в рамках БССС. Это то, что касается синтаксиса естественного языка. Кстати, эта платформа структурной организации языка в качестве «базового синтаксиса» была презентована автором еще на конференции в Варне по компьютеризации естественных языков [21].

Семантика. Относительно текущего состояния семантики отсылаем читателей снова к «Первому московскому лингвистическому альманаху», где один из фундаторов модели «Смысл – Текст» Н.В. Перцов дает непредубежденную характеристику существующих

семантических концепций, и оценка эта не очень утешительная. Во всех семантических построениях объектом исследования выбирается слово (узел), а не знание (не структура, связывающая ряд категорий определенными функциональными зависимостями); важным является также замечание, что «*семантика слова*» представляется лишь бледной тенью «*семантики образа*» [15]. По нашему предположению семантика (как совокупность знаний) должна формироваться в базе знаний при формировании модели мира посредством взаимосвязанного множества отдельных квантов знаний. В концентрированном виде новый взгляд на структурную организацию языка представлен работами [22], [23].

Предложение. Если теперь вернуться к проблемам информационных технологий, можем констатировать, что предложенное системное видение структурной организации языка во многом с иных позиций представляет проблемные вопросы информационного поиска. Если еще вспомним, что информационный поиск пытается моделировать одну из составляющих речевой деятельности человека – обращение к знаниям (смотри модель ИРС), то уже становится очевидным, что данный подход существенно изменяет идеологию формирования лингвистического процессора и, соответственно – его модели M_1, а также выдает нам определенные ориентиры для формирования базы знаний в виде M_2, где элементами восприятия, хранения и накопления информации представляются уже отдельные кванты/фрагменты знаний, а не совершенно неструктурированный текст. Поиск, формируемый при таких условиях, определим как *«поиск по знаниям»* в отличие от существующей идеологии *«поиска по ключевым словам»*. Практически теперь мы имеем новую информационную платформу для моделирования всего кластера информационных естественно-языковых технологий.

2 Эмуляция «поиска по знаниям»

Модель поиска «по знаниям». Сложность моделирования предложенной схемы поиска определяется следующими особенностями. Все существующие поисковики, практически, ориентированы на поиск по ключевым словам (отдельным или их множествам, связанным или нет логическими отношениями И, ИЛИ, НЕ). В последнее время поиск иногда ведется уже по полному тексту, однако, при этом абсолютно игнорируется внутренняя структура сообщения. Известно, что любое предложение как структура представляется множеством категорий, связанных определенными функциональными зависимостями; ни первое, ни второе не учитывается современными поисковиками; не используются также предложенные еще Н. Хомским [24] трансформационные грамматики; кроме того, от пользователя скрыты особенности индексирования исходных массивов и нюансы собственно процедуры поиска. В соответствии с таким положением вещей сложно сравнивать эффективность работы различных поисковиков – конечная оценка остается всегда за пользователями. Все эти факторы определяют сложность реализации и оценки предложенного подхода к поиску на фоне развитой структуры WEB. Тем не менее, процедура проверки предложенного подхода с целью сравнения с существующими технологиями представляется следующей.

Если мы определяем «знание» (или его часть) как некоторое множество категорий, связанных определенными функциональными отношениями, то традиционный поиск по ключевым словам необходимо теперь заменить поиском по структурам с идентификацией категорий и функций. Это определенным образом приближает нас к поиску по полному тексту. Особенность лишь в том, что человек однозначно идентифицирует (понимает) текст независимо от всех его трансформаций либо от порядка слов, что особенно важно для информационных технологий, ориентированных на обработку флективных языков. В последнем случае даже отдельная базовая структура как

«квант знаний» может быть развернута более сорока миллиардами вариантов, что сразу ставит под сомнение процедуру поиска по полному тексту; обойти это препятствие возможно лишь при дополнительных ограничениях и преобразованиях. Человек эффективно решает эту проблему понимания за счет обращения к своим знаниям и широким возможностям их трансформирования; современные же технологии пока не обладают возможностью такой гибкой трансформации, как и не обладают вообще некоторым аналогом памяти человека – базой знаний. Тем не менее, покажем, насколько эффективным представляется дополнение существующих процедур поиска лишь некоторыми возможностями учета структур запроса и их трансформирования.

Эксперимент. Итак, суть виртуального эксперимента по модифицированию информационного поиска представляется следующим образом: выбирается произвольный запрос и обрабатывается на пяти ведущих поисковиках; двадцать первых документов выдачи (презумпция ранжирования по релевантности ???) каждого поисковика выбираются в качестве тестовых массивов для реализации «поиска по знаниям». Поиск на тестовых массивах выполняется вручную экспертом, который, конечно же, при определении релевантности и трансформирования структур пользуется своими «знаниями». При этом авторы использовали при виртуальном поиске лишь три из всего спектра перечисленных ранее трансформаций: исходный запрос с глагольной формой предиката трансформировался последовательно в субстантивную, адъективную и адвербиальную формы. В тестовых массивах для каждого поисковика вручную выполнялся поиск одновременно по трем трансформированным структурам и за определенным критерием определялась релевантность документов выдачи.

Объектом исследования, таким образом, является интернет – система, более точно – технология поиска информации. Целью виртуального эксперимента было сравнить эффективность существующих схем поиска (поиск по ключевым словам) с предлагаемой схемой поиска по структурам (поиск по знаниям).

Тестированию подверглись пять ведущих систем:

– Google – наиболее популярная поисковая система (http//www.google.com/);
– Yahoo – одна из ведущих американских поисковых систем (http//yahoo.com/);
– Bing – популярная система корпорации Microsoft (http//www.bing.com/);
– Yandex – самая известная российская поисковая система (http//www.yandex.ru/);
– Мета – украинская поисковая система (http//meta.ua/).

В качестве запроса выбрана была следующая фраза: *«Писатель, который получил Пулитцеровскую премию в 2009 году»*.

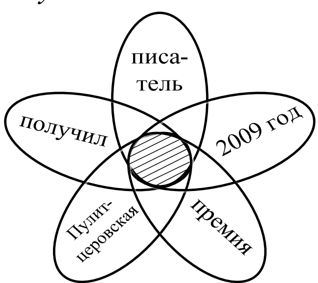


Рисунок 1 – Графическое представление запроса Q_1

Для поиска по ключевым словам эта информационная потребность представляется в виде последовательности: $Q_1 = ($ *писатель* $\cap$ *получил* $\cap$ *Пулитцеровская* $\cap$ *премия* $\cap$ *«2009год»*$)$, где $\cap$ означает логическое «И». Графическая интерпретация такого запроса представлена рис. 1, где заштрихованная область отображает конъюнкцию ключевых слов запроса.

В таком виде запрос обрабатывается на всех поисковиках; при этом по каждому из них фиксируется выдача первых двадцати документов, на которых в дальнейшем вручную выполняется поиск «по знаниям».

Для виртуального (ручного) поиска исходный запрос представляется в первозданном виде, зафиксированном на уровне базовых структур (рис. 2).

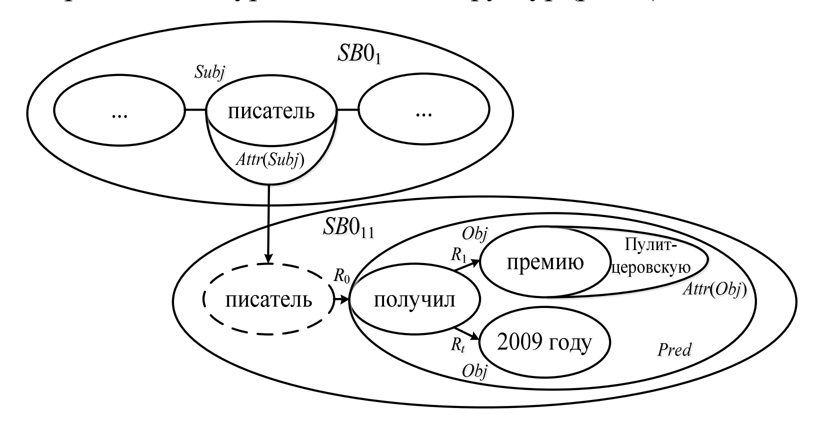


Рисунок 2 – Структура запроса Q_2 при поиске по знаниям

Категории запроса: $Subj$ – субъект, $Attr(Subj)$ – атрибут субъекта, $Pred$ – предикатор, Obj – объект, $Attr(Obj)$ – атрибут объекта, функции запроса: R_0 – иметь предикат, R_1 – иметь объект, R_t – иметь временную характеристику.

Здесь мы имеем структуру естественно-языкового сообщения, реализованную на множестве категорий «писатель», «получил», «премию», «Пулитцеровскую», «2009 год», связанных определенными отношениями R_0, R_1, R_t. Именно совокупность категорий и связывающих их функций однозначно определяет смысл (семантику) сообщения. Указанная структура уже как полнотекстовая на смысловом уровне выбирается в качестве запроса для виртуального поиска на полученных выборках из двадцати документов для каждого поисковика. Кроме того, исходя из авторского видения структурной организации текстовой информации, семантически адекватными запросу Q_2 будут и сообщения, фрагменты которых представлены его трансформациями. Таким образом, «поиск по знаниям» одновременно включает в себя поиск по всем разновидностям запроса:

– Q_2 – Писатель, который получил Пулитцеровскую премию в 2009 году . . . . ,
– Q_{21} – Писатель, получивший Пулитцеровскую премию в 2009 году . . . . ,
– Q_{22} – Писатель, получая Пулитцеровскую премию в 2009 году . . . . ,
– Q_{23} – Получение писателем Пулитцеровской премии в 2009 году . . . .

Важно здесь обратить внимание на многоточие в конце каждой разновидности запроса – это означает, что релевантными будут документы, сообщения которых включают в себя как минимум указанные разновидности структур Q_2, Q_{21}, Q_{22}, Q_{23}.

Для сравнения эффективности поиска по двум стратегиям воспользуемся оценкой релевантности полученных результатов. При поиске по ключевым словам запрос Q_1 – соответствие документа выдачи информационной потребности пользователя (его релевантность) оценивалась следующим образом. Если в документе выдачи в одном сообщении совпали пять ключевых слов, то выставлялась оценка «пять», если четыре, то оценка – «четыре» и т.д. до единицы. Полученные данные отображаются в виде графиков для каждой поисковой системы (рис. 3*а*).

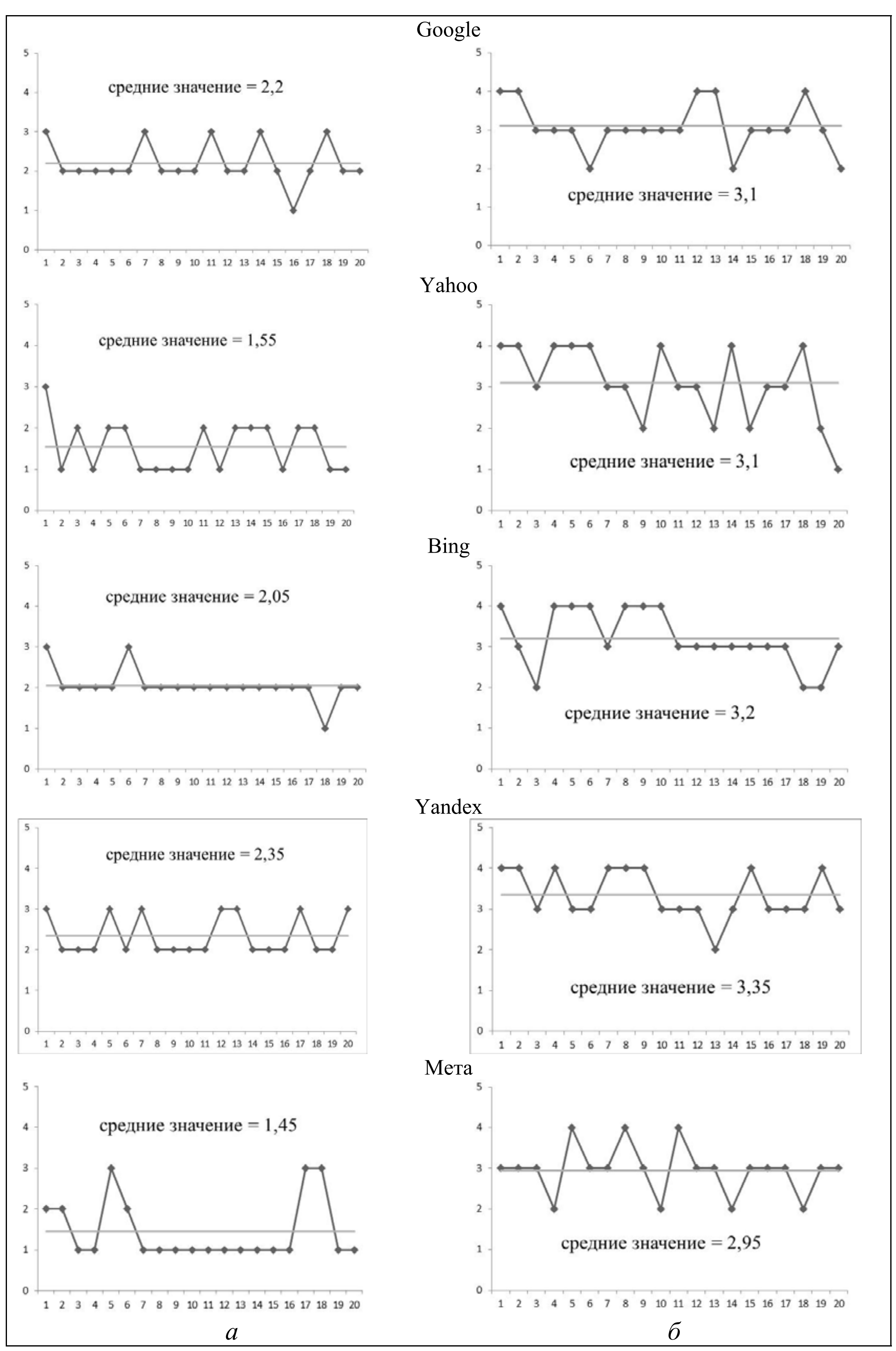


Рисунок 3 – Сравнительная характеристика оценивания релевантности (*а* – оценка релевантности с помощью поиска по ключевым словам; *б* – оценка релевантности с помощью поиска по знаниям)

Во втором случае – «поиск по знаниям» – сравниваются уже фрагменты структур (как совокупность категорий, связанных определенными отношениями). Совпадение требует идентификации не только категорий, но и функций, их связывающих. Таким образом, если в документе (отдельных его сообщениях) имеется структура, совпадающая целиком со структурой запроса из пяти элементов с учетом их функциональных связей, то выставляется оценка «пять», если совпадает четыре элемента совместно с их функциональными связями, то выставляется оценка «четыре» и т. д. Результаты «поиска по знаниям» для всех тестируемых поисковиков (оценка релевантности) представлены рис. 3*б)*.

При сравнении результатов сразу становится очевидным значительное повышение релевантности выдачи «поиска по знаниям» для всех поисковиков: Google (+0,9), Yahoo (+1,55), Bing (+1,15), Yandex (+1,0), Мета (+1,5). Конечно, эти результаты получены на очень коротких выборках и, вероятно, заранее были прогнозируемы; однако следует подчеркнуть очень существенные обстоятельства: во-первых, используемые при поиске «по знаниям» трансформации вполне доступны для реализации на поисковиках по чисто формальным критериям, во-вторых, здесь решается важный принципиальный момент – мы получаем гарантию того, что после проведения поиска в подобном режиме мы не пропустим интересующих нас релевантных документов, в-третьих, мы можем значительно более точно в запросе формулировать свою информационную потребность, поскольку сравнение ведется «по знаниям».

Заключительная часть. Таким образом, мы получили в итоге результат сравнения машинного поиска с поиском информации человеком. Приоритет, конечно же, остается за человеком, и такой исход можно было бы предположить заранее, учитывая что:

– машина до сих пор не снабжена знаниями, в которых зафиксирована модель нашей среды обитания как реальной, так и виртуальной;

– не имеет достаточного опыта общения со средой;

– не обладает интуицией, предвидением и многими-многими другими способностями;

– не обладает той языковой компетенцией, которая спрессована в современных грамматиках лингвистических фолиантов и лишь очень незначительно учитывается в современных информационных технологиях.

Резюмируя, можем подчеркнуть еще раз то, о чем речь шла ранее: *до сих пор мы имеем лишь очень приблизительное представление о самом объекте исследования – индивидуальной речевой системе человека, реализующей речевую деятельность в режимах синтеза/анализа языкового сообщения, и в современные информационные технологии заложена лишь очень незначительная и несистемная толика наших знаний об этой – одной из наиболее сложных интеллектуальных функций человека.*

В эксперименте же были учтены лишь два момента языковой организации: отдельные схемы трансформирования сообщения наподобие порождающих грамматик Н. Хомского, дополненные еще особенностями структурной организации языка, рассматривающими произвольный текст как множество однотипных синтаксических структур. И первое, и второе требуют лишь ряда формализованных процедур, касающихся трансформации как запроса, так и текстов документов. Однако на сегодняшний день при современном развитии вычислительной техники и систем программирования – это, по существу, рутинные стандартные процедуры, которые без труда могут быть переложены на плечи компьютера.

Заключительный тезис фактически повторяет изначальное положение вводной части: чтобы более-менее адекватно моделировать отдельные аспекты речевой деятельности человека (и поиск информации в том числе) необходимо хорошо изучить сам объект

исследования – индивидуальную речевую систему человека как совокупность лингвистического процессора и базы знаний – построить затем модели электронные обеих составляющих и замкнуть их друг на друга. Это практически единственный путь совершенствования и развития всех ЕЯ-технологий. Однако отдельные частные вопросы можно и нужно решать уже сегодня, но на принципиально другой основе, формируемой на стыке всех научных направлений, связанных с исследованием речевой деятельности человека.

Литература

1. Шишков Д.П. Компьютеризация естественных языков / Д.П. Шишков // Компьютеризация естественных языков : труды первого Международного семинара. – Болгария, Варна, курорт «Св. Константина», 3 – 7 сентября, 1999. – С. 31-42.
2. Речицкий В.В. Символическая реальность и право / Речицкий В.В. – Львов : ВНТЛ-Классика, 2007. – 732 с.
3. Кисленко Ю.І. Інформаційний підхід до аналізу структурного рівня мовної організації / Ю.І. Кисленко // Штучний інтелект. – 2010. – № 4. – С. 90-101.
4. Кисленко Ю.И. Перспективы развития естественно-языковых технологий / Ю.И. Кисленко // Системні дослідження та інформаційні технології. Інститут прикладного системного аналізу НАН України та Міносвіти і науки. – 2004. – № 2.
5. Поляков В.Н. Интеллектуальная поисковая машина. Концептуальный проект / В.Н. Поляков // Труды Казанской школы по компьютерной и когнитивной лингвистике. TEL-2000 (Казань, 17 – 20 октября 2000). – Казань : Изд-во Сэлэт, 2000. – Вып. 5.
6. Поиск 2.0, каким он, возможно, будет [Электронный ресурс]. – Режим доступа : http://habrahabr.ru/blogs/search_engines/94426/
7. Ландэ Д.В. Поиск знаний в INTERNET / Ландэ Д.В. – Изд-во «Диалектика-Вильямс». – 2005.
8. Ландэ Д.В. От идеи к технологии / Д.В. Ландэ // ТЕЛЕКОМ. – 2005. – № 6.
9. Covles P. Web Services and the Semantic Web / P. Covles // Web Services Journal. – December 2002. – Vol. 02. – Is. 12.
10. Berners-Lee T. The Semantic Web / T. Berners-Lee, J. Hendler, O. Lassila // Scientific American. – May 17, 2001.
11. Поляков В.Н. Использование технологий, ориентированных на лексическое знание / В.Н. Поляков // Проблемы прикладной лингвистики. – 2003.
12. Осипов Г.С. Интеллектуальный поиск в глобальных и локальных вычислительных сетях и базах данных / Г.С. Осипов, И.А. Тихомиров, И.В. Смирнов // Программные системы: теория и приложения : труды Международной конференции, (Переславль-Залесский). – М. : Физматлит, 2004. – Т. 2. – С. 21-34.
13. Щерба Л.В. О трояком аспекте языковых явлений и эксперименте в языкознании / Щерба Л.В. // Щерба Л.В. Языковая система и речевая деятельность. – М., 1974.
14. Мельников Г. П. Системология и языковые проблемы кибернетики / Мельников Г. П. – М. : Сов. Радио, 1978. – 368 с.
15. Перцов Н.В. О некоторых проблемах современной семантики и компьютерной лингвистики / Н.В. Перцов // Московский лингвистический альманах. – 1996. – Вып. 1. – С. 9-66.
16. Грамматика современного русского литературного языка (Грамматика –70). – М. : Наука, 1970.
17. Астахова Л.И. Предложение и его членение (прагматика, семантика, синтаксис) / Астахова Л.И. – Днепропетровский ГУ, 1992.
18. Компьютерная лингвистика: моделирование языкового общения / [пер. с англ., сост., ред. и вступ. ст. Б.Ю. Городецкого]. – М.: Прогресс,1989. – Вып. 24.– 432 с. – (Серия «Новое в зарубежной лингвистике»)
19. Кисленко Ю.І. Нейрофізіологічне підґрунтя структурної організації мовного матеріалу / Ю.І. Кисленко // Доповіді НАН України. – 2007. – № 11.– С. 158-164.
20. Кисленко Ю. От мысли к знанию (нейрофизиологические основания) : монографія / Кисленко Ю. – К. : Издательство «Український літопис». 2008. – 101 с.
21. Кисленко Ю.И. Базовый синтаксис / Ю.И. Кисленко // Компьютеризация естественных языков : труды первого Международного семинара. – Болгария, Варна, курорт «Св. Константина», 3 – 7 сентября, 1999. – С. 82-92.
22. Кисленко Ю.І. Системна організація мови : монографія / Кисленко Ю.І. – К. :Український літопис, 1997. – 217 с.
23. Кисленко Ю.І. Архітектура мови (лінгвістичне забезпечення інтелектуальних інтегрованих систем) : Учбовий посібник / Кисленко Ю.І. – К. : Віпол, 1998. – 343 с.
24. Хомский Н. Аспекты теории синтаксиса / Хомский Н. – М. : МГУ, 1972 .

Literatura

1. Shishkov D.P. Trudy pervogo Mezhdunarodnogo seminara "Komp'juterizacija estestvennyh jazykov". Bolgarija, Varna, kurort "Sv. Konstantina". 3 – 7 sentjabrja. 1999. S. 31-42.
2. Rechickij V.V. Simvolicheskaja real'nost' i pravo. L'vov : VNTL-Klassika. 2007. 732 s.
3. Kislenko Ju.I. Shtuchnij intelekt. 2010. № 4. S. 90-101.
4. Kislenko Ju. I. Sistemni doslidzhennja ta informacijni tehnologiï. Institut prikladnogo sistemnogo analizu NAN Ukraïni ta Minosviti i nauki. 2004. №. 2.
5. Poljakov V.N. Trudy Kazanskoj shkoly po komp'juternoj i kognitivnoj lingvistike. TEL-2000. Kazan'. 17-20 oktjabrja. Vyp. 5. 2000.
6. Poisk 2.0, kakim on, vozmozhno, budet. http://habrahabr.ru/blogs/search_engines/94426/
7. Landje D.V. Poisk znanij v INTERNET. Izd-vo "Dialektika-Vil'jam". 2005.
8. Landje D.V. Ot idei k tehnologii. TELEKOM. № 6. 2005.
9. Covles P. Web Services Journal. December 2002. Vol. 02. Issue 12.
10. Berners-Lee T. Scientific American. May 17, 2001.
11. Poljakov V.N. Ispol'zovanie tehnologij, orientirovannyh na leksicheskoe znanie. Problemy prikladnoj lingvistiki. 2003.
12. Osipov G.S. Trudy mezhdunarodnoj konferencii "Programmnye sistemy: teorija i prilozhenija". Pereslavl'-Zalesskij, M.: Fizmatlit. 2004. T. 2. S. 21-34.
13. Shherba L.V. O trojakom aspekte jazykovyh javlenij i jeksperimente v jazykoznanii. Jazykovaja sistema i rechevaja dejatel'nost'. M. 1974.
14. Mel'nikov G. P. Sistemologija i jazykovye problemy kibernetiki. M.: Sov. Radio. 1978. 368 s.
15. Percov N. V. Moskovskij lingvisticheskij al'manah. Vyp 1. 1996. S. 9-66.
16. Grammatika sovremennogo russkogo literaturnogo jazyka (Grammatika-70). M. : Nauka. 1970.
17. Astahova L.I. Predlozhenie i ego chlenenie (pragmatika, semantika, sintaksis). Dnepropetrovskij GU. 1992.
18. Komp'juternaja lingvistika: modelirovanie jazykovogo obshhenija (per. s angl., sost., red. i vstup. St. B.Ju. Gorodeckogo). "Serija Novoe v zarubezhnoj lingvistike". M.: Progress. Vyp 24. 1989. 432 s.
19. Kislenko Ju.I. Dopovidi NAN Ukraïni. № 11. 2007. K. : Vidavnichij dim "Akademperiodika". S. 158-164.
20. Kislenko Ju. Ot mysli k znaniju (nejrofiziologicheskie osnovanija): monografija. K.: Izdatel'stvo "Ukraïns'kij litopis". 2008. 101 s.
21. Kislenko Ju.I. Trudy pervogo Mezhdunarodnogo seminara "Komp'juterizacija estestvennyh jazykov". Bolgarija, Varna, kurort "Sv. Konstantina". 3 – 7 sentjabrja. 1999. S. 82-92.
22. Kislenko Ju. I. Sistemna organizacija movi: Monografija. K. : Ukraïns'kij litopis. 1997. 217 s.
23. Kislenko Ju.I. Arhitektura movi (lingvistichne zabezpechennja intelektual'nih integrovanih sistem) : Uchbovij povibnik. K. : Vipol. 1998. 343 s.
24. Homskij N. Aspekty teorii sintaksisa. M. : MGU, 1972.

Ю.І. Кисленко, А.В. Терентьєв

Проблеми та перспективи розвитку пошукових систем

Розглянуті важливі проблеми функціонування провідних пошукових систем – недостатній облік та використання природної мови на етапах формування запиту та пошуку інформації. Перспектива бачиться у використанні нового підходу до структурної організації мови, який знімає багато суперечностей класичної лінгвістики та який дозволяє замінити «пошук за ключовим словом» на «пошук за знаннями».

Yu.I. Kislenko, A.V. Terentiev

Problems and Prospects of Search Engines

The not seen at once problem of the leading search engines is low consideration and application of natural language on the stages of demand assignment and information search. The prospects are considered in application of a new approach to the structural organization of language. This approach reduces contradiction of classical linguistics and delivers to change "keyword search" to "knowledge search".

Статья поступила в редакцию 16.06.2011.